Brain Computer Interaction

Feature Translation

Course Instructors

Dr. Annushree Bablani

Acknowledgements: Dr. Sreeja S R

THIRD STEP: FEATURE CONDITIONING

- The distributions and the relationships among the features can have a significant effect on the performance of the translation algorithm that follows feature extraction. These effects depend on the characteristics of the particular translation algorithm.

  - ***NORMALIZATION***
  - ***LOG-NORMAL TRANSFORMS***
  - ***FEATURE SMOOTHING***
  - ***PCA AND ICA***
  - ***Removing irrelevant and redundant features (MRMR-Maximum Relevant Minimum Redundant)***

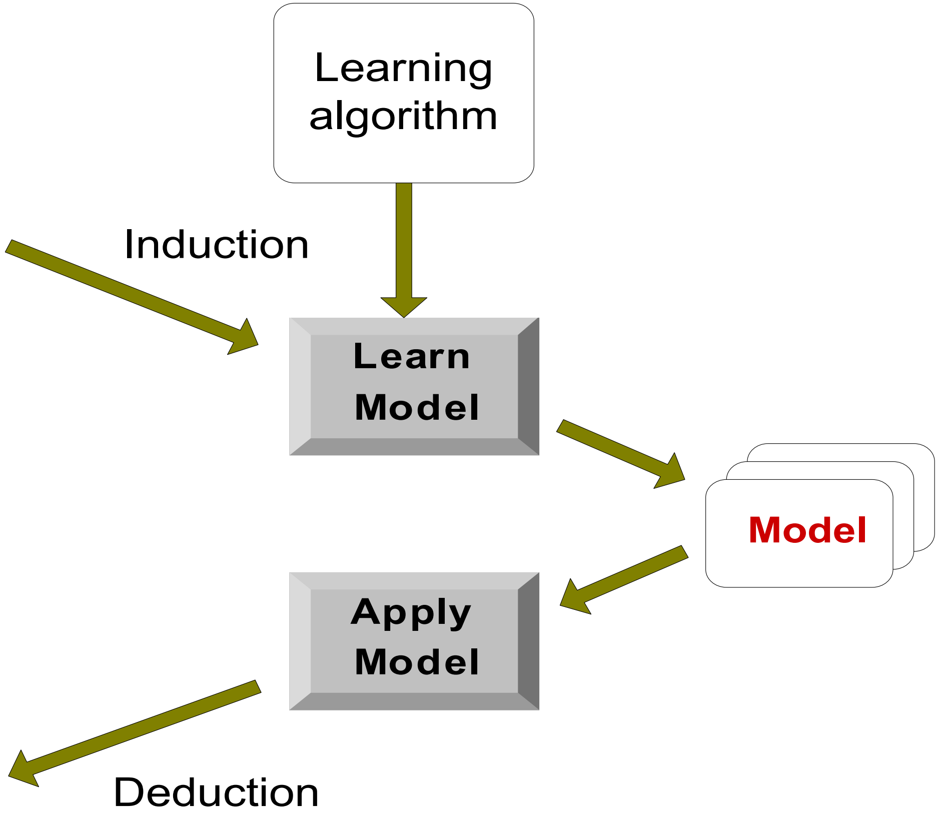
Feature Translation

- Discriminant functions
- Regression functions

Illustrating Classification Tasks

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



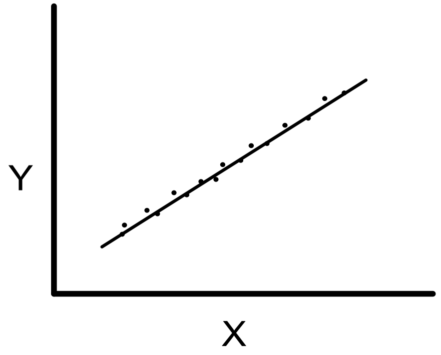
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

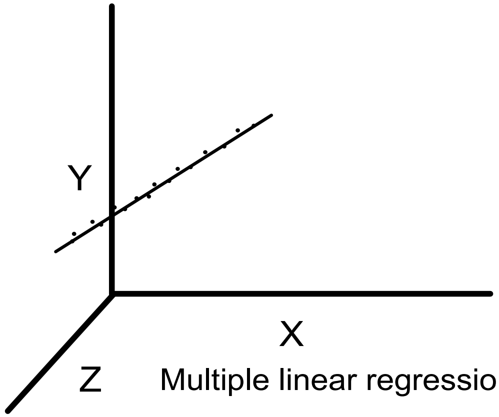
Regression Analysis

- The regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables, which can be used for the purpose of prediction of the values of dependent variable, given the values of independent variables.

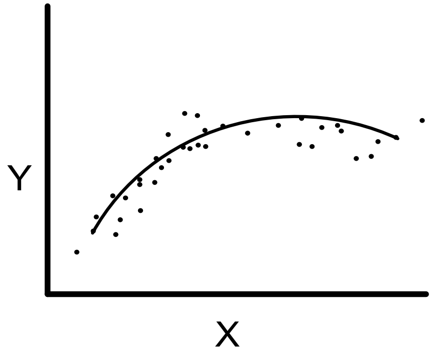
- **Classification of Regression Analysis Models**
  - Linear regression models
    1. Simple linear regression
    2. Multiple linear regression
  - Non-linear regression models



Simple linear regression



Multiple linear regression

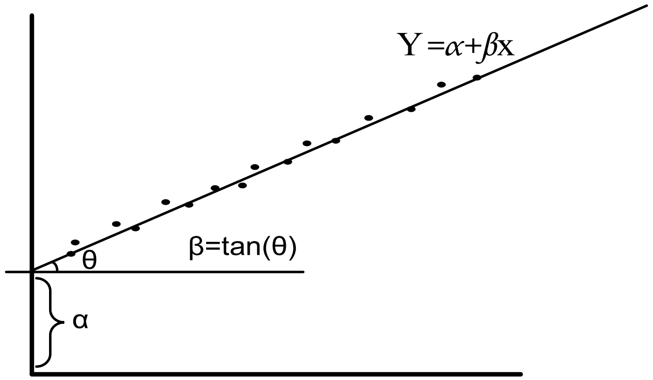


Non-linear regression

Simple Linear Regression Model

In simple linear regression, we have only two variables:

- Dependent variable (also called Response), usually denoted as .
- Independent variable (alternatively called Regressor), usually denoted as .
- A reasonable form of a relationship between the Response and the Regressor is the linear relationship, that is in the form

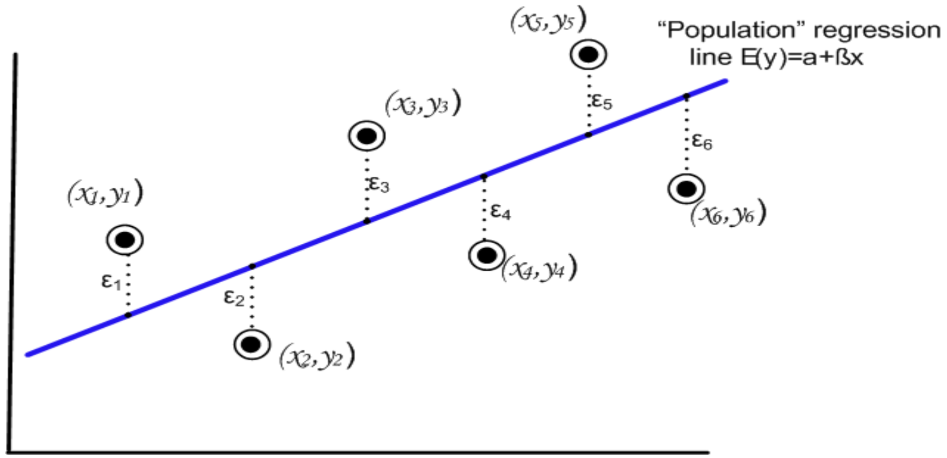


Note:

- There are infinite number of lines (and hence)

- The concept of regression analysis deal with finding the best relationship between and (and hence best fitted values of) quantifying the strength of that relationship.

Regression Analysis



Given the set of data involving pairs of values, our objective is to find "true" or population regression line such that

Here, is a random variable with and . The quantity is often called the **error variance**.

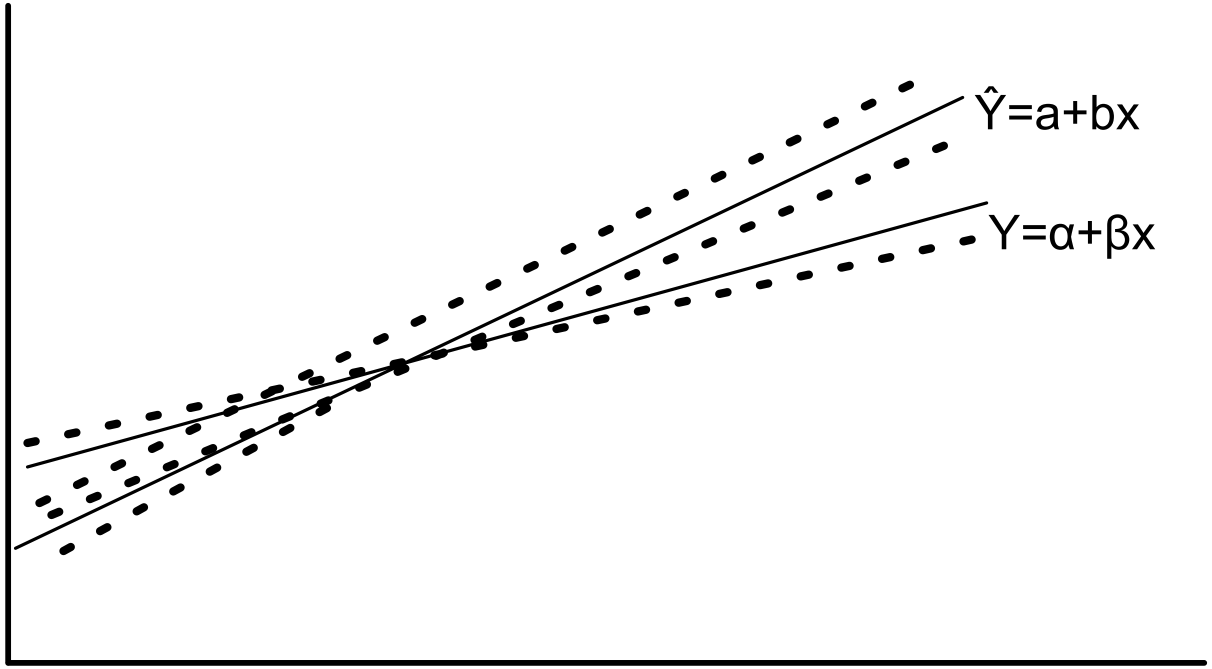
Note:

- implies that at a specific , the values are distributed around the "true" regression line (i.e., the positive and negative errors around the true line is reasonable).
- are called **regression coefficients**.
- values are to be estimated from the data.

True versus Fitted Regression Line

- The task in regression analysis is to estimate the regression coefficients .
- Suppose, we denote the estimates *a* for and *b* for . Then the fitted regression line is

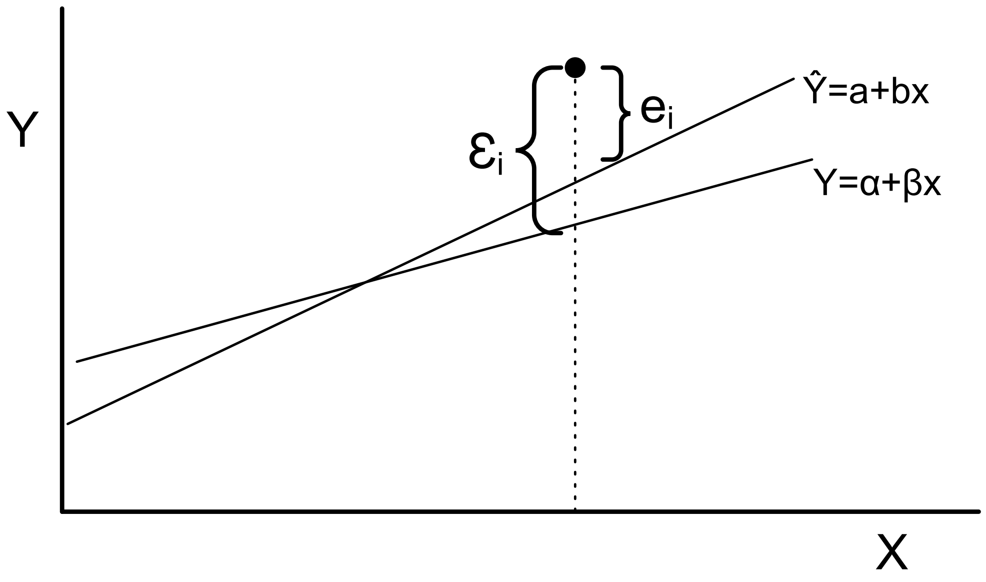
where is the predicted or fitted value.



Least Square Method to estimate

This method uses the concept of residual. A residual is essentially an error in the fit of the model . Thus, residual is

,



Least Square method

- The residual sum of squares is often called **the sum of squares of the errors** about the fitted line and is denoted as SSE

$$\text{SSE} = \quad =$$

- We are to minimize the value of SSE and hence to determine the parameters of *a* and *b*.

- Differentiating SSE with respect to *a* and *b*, we have

For minimum value of SSE, $\quad 0$

$$0$$

Least Square method to estimate

Thus we set

+b=

These two equations can be solved to determine the values of and *b*, and it can be calculated that

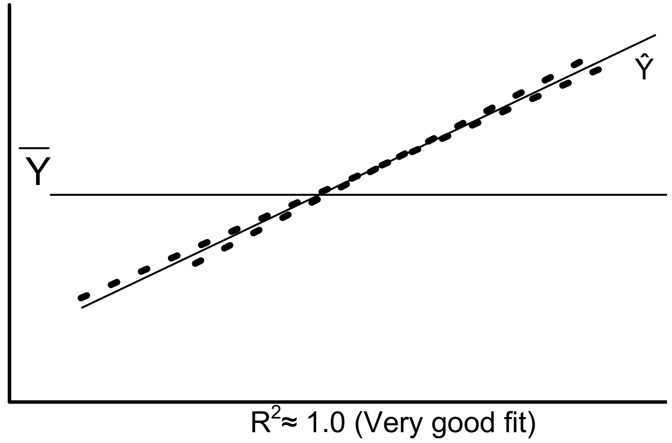
: Measure of Quality of Fit

- A quantity , is called **coefficient of determination** is used to measure the proportion of variability of the fitted model.
- We have
- It signifies the **variability due to error**.
- Now, let us define the total corrected sum of squares, defined as
- SST represents the variation in the response values. The is

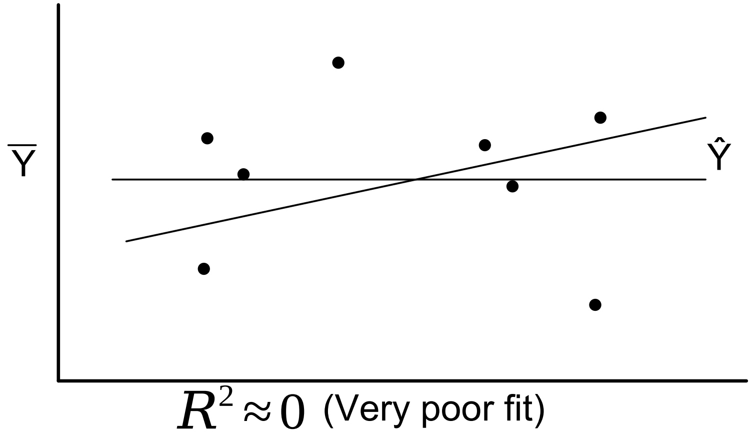
Note:

- If fit is perfect, all residuals are zero and thus = 1.0 (very good fit)
- If SSE is only slightly smaller than SST, then (very poor fit)

: Measure of Quality of Fit



$R^2 \approx 1.0$ (Very good fit)

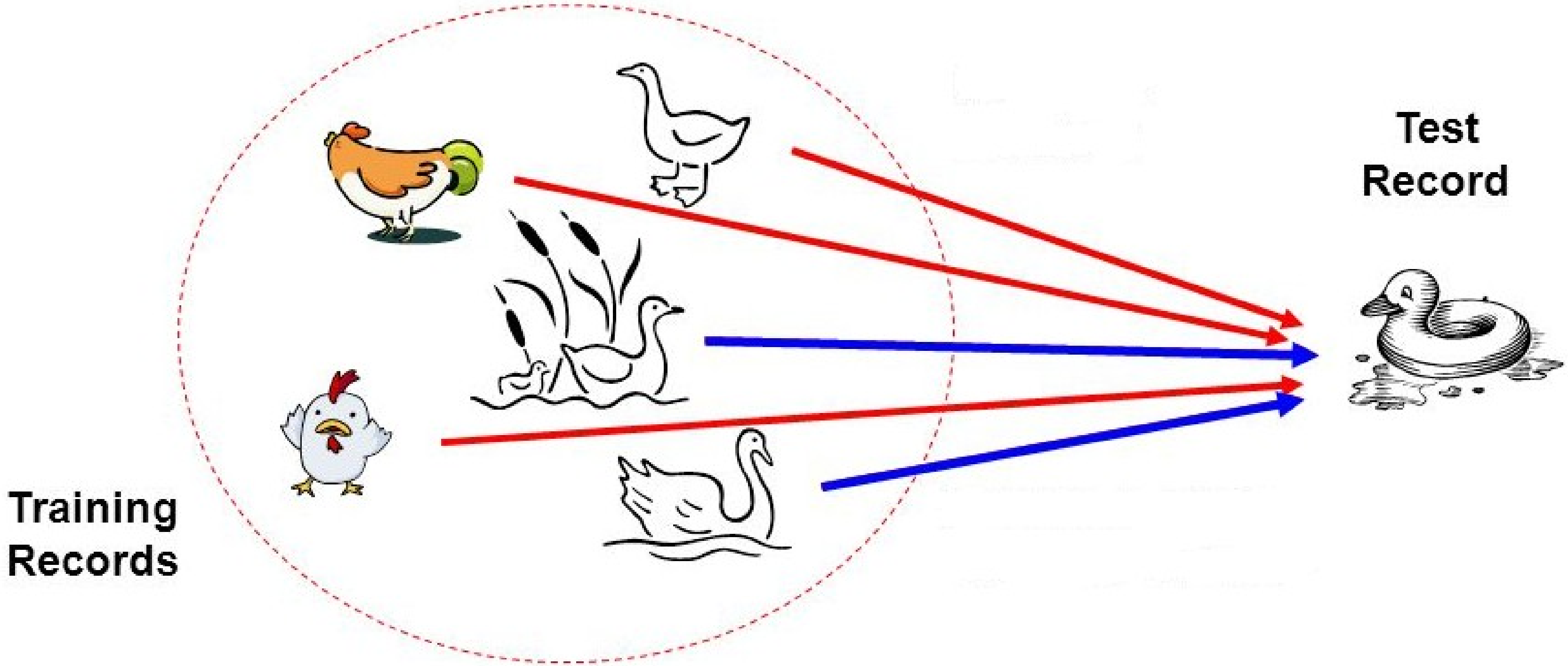


$R^2 \approx 0$ (Very poor fit)

Bayesian Classifier

Bayesian Classifier

- Principle
  - If it walks like a duck, quacks like a duck, then it is probably a duck



Bayesian Classifier

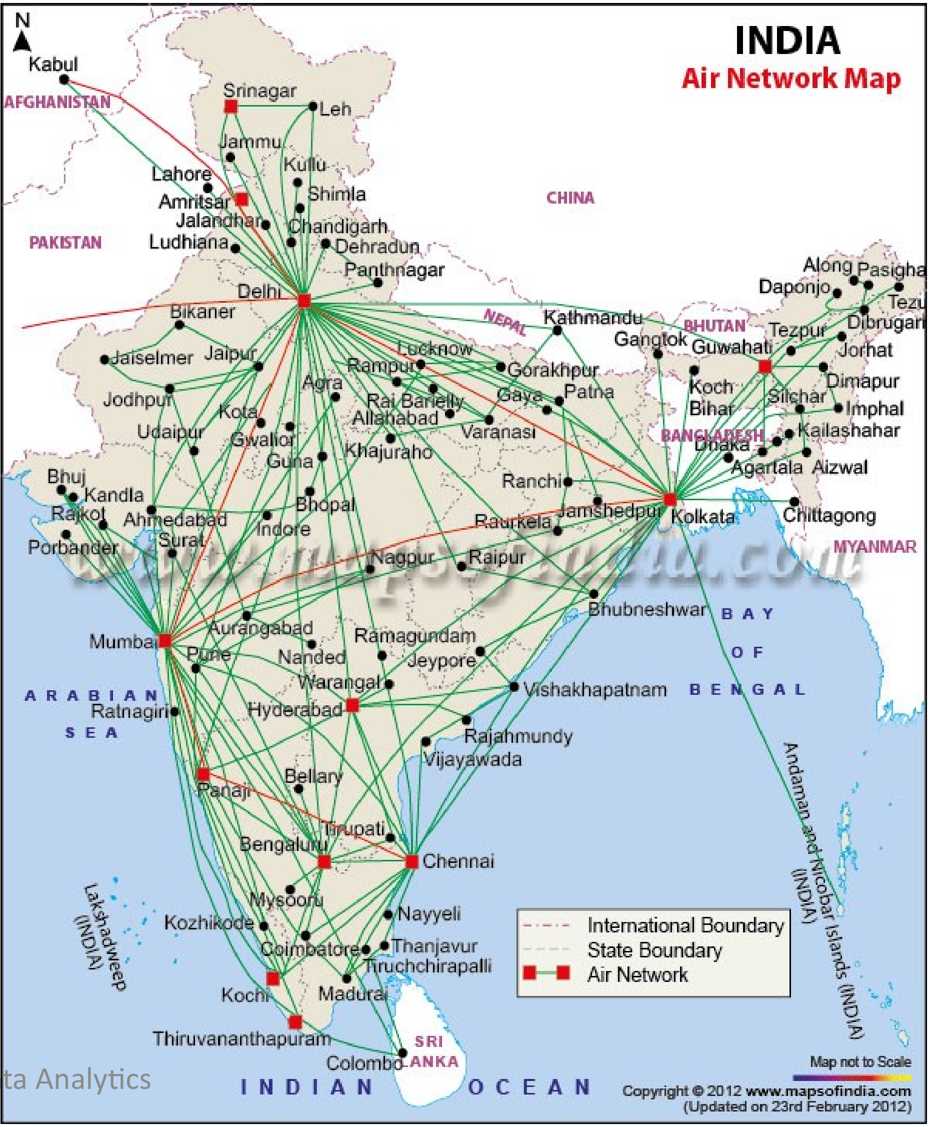
- A statistical classifier
  - Performs *probabilistic prediction*, *i.e.,* predicts class membership probabilities
- Foundation
  - Based on Bayes' Theorem.
- Assumptions

  1.The classes are mutually exclusive and exhaustive.

  2.The attributes are independent given the class.
- Called "Naïve" classifier because of these assumptions.
  - Empirically proven to be useful.
  - Scales very well.

Example: Bayesian Classification

- **Example 8.2:** Air Traffic Data
  - Let us consider a set observation recorded in a database
    - Regarding the arrival of airplanes in the routes from any airport to New Delhi under certain conditions.



Air-Traffic Data

Days	Season	Fog	Rain	Class

Cond. to next slide…

Air-Traffic Data

Cond. from previous slide…

Days	Season	Fog	Rain	Class

Air-Traffic Data

- In this database, there are four attributes

  A = [Day, Season, Fog, Rain]

  with 20 tuples.
- The categories of classes are:

  C= [On Time, Late, Very Late, Cancelled]

- Given this is the knowledge of data and classes, we are to find most likely classification for any other unseen instance, for example:

Week Day	**Winter**	**High**	**None**	**???**

- Classification technique eventually to map this tuple into an accurate class.

Bayesian Classifier

- In many applications, the relationship between the attributes set and the class variable is non-deterministic.
  - In other words, a test cannot be classified to a class label with certainty.
  - In such a situation, the classification can be achieved probabilistically.
- The Bayesian classifier is an approach for modelling probabilistic relationships between the attribute set and the class variable.
- More precisely, Bayesian classifier use Bayes' Theorem of Probability for classification.
- Before going to discuss the Bayesian classifier, we should have a quick look at the Theory of Probability and then Bayes' Theorem.

Bayes’ Theorem of Probability

Simple Probability

Definition 8.2: **Simple Probability**

If there are n elementary events associated with a random experiment and m of n of them are favorable to an event A, then the probability of happening or occurrence of A is

Simple Probability

- Suppose, A and B are any two events and *P(A)*, *P(B)* denote the probabilities that the events *A* and *B* will occur, respectively.

- **Mutually Exclusive Events:**
  - Two events are mutually exclusive, if the occurrence of one precludes the occurrence of the other.

Example: Tossing a coin (two events)

Tossing a ludo cube (Six events)

💡 Can you give an example, so that two events are not mutually exclusive?

Hint: Tossing two identical coins, Weather (sunny, foggy, warm)

Simple Probability

- **Independent events:** Two events are independent if occurrences of one does not alter the occurrence of other.

Example: Tossing both coin and ludo cube together.

(How many events are here?)

- Can you give an example, where an event is dependent on one or more other events(s)?

Hint: Receiving a message (A) through a communication channel (B) over a computer (C), rain and train.

Joint Probability

Definition 8.3: **Joint Probability**

If *P(A)* and *P(B)* are the probability of two events, then

If *A* and *B* are mutually exclusive, then
If *A* and *B* are independent events, then

Thus, for mutually exclusive events

Conditional Probability

Definition 8.2: **Conditional Probability**

If events are dependent, then their probability is expressed by conditional probability. The probability that *A* occurs given that *B* is denoted by .

Suppose, *A* and *B* are two events associated with a random experiment. The probability of *A* under the condition that *B* has already occurred and is given by

Conditional Probability

Corollary 8.1: **Conditional Probability**

or

For three events A, B and C

For n events $A_1, A_2, \ldots, A_n$ and if all events are mutually independent to each other

Note:

if events are **mutually exclusive**

if A and B are **independent**

otherwise**,**

Conditional Probability

- Generalization of Conditional Probability:

$$\because \ P(A) = P(B)$$

By the law of total probability : P(B) =

Conditional Probability

In general,

Total Probability

Definition 8.3: **Total Probability**

Let be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with , then

Total Probability: An Example

Example 8.3

A bag contains 4 red and 3 black balls. A second bag contains 2 red and 4 black balls. One bag is selected at random. From the selected bag, one ball is drawn. What is the probability that the ball drawn is red?

This problem can be answered using the concept of Total Probability

Selecting bag *I*

Selecting bag *II*

A = Drawing the red ball

Thus,

where, = Probability of drawing red ball when first bag has been chosen

and = Probability of drawing red ball when second bag has been chosen

Reverse Probability

Example 8.3:

A bag (Bag I) contains 4 red and 3 black balls. A second bag (Bag II) contains 2 red and 4 black balls. You have chosen one ball at random. It is found as red ball. What is the probability that the ball is chosen from Bag I?

Here,

Selecting bag *I*

Selecting bag *II*

A = Drawing the red ball

We are to determine P(|A). Such a problem can be solved using Bayes' theorem of probability.

Bayes' Theorem

Theorem 8.4: **Bayes' Theorem**

Let be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with , then

Prior and Posterior Probabilities

- P(A) and P(B) are called prior probabilities
- P(A|B), P(B|A) are called posterior probabilities

Example 8.6: Prior versus Posterior Probabilities

- This table shows that the event *Y* has two outcomes namely *A* and *B*, which is dependent on another event *X* with various outcomes like and .
- **Case1:** Suppose, we don't have any information of the event *A*. Then, from the given sample space, we can calculate *P(Y = A)* = = 0.5

.

- **Case2:** Now, suppose, we want to calculate *P(X = |Y =A)* = = 0.4 .

The later is the conditional or posterior probability, where as the former is the prior probability.

X	*Y*
	A
	A
	B
	A
	B
	A
	B
	B
	B
	A

Naïve Bayesian Classifier

- Suppose, *Y* is a class variable and *X* = is a set of attributes,

  with instance of *Y*.

INPUT (X)	CLASS(Y)

- The classification problem, then can be expressed as the class-conditional probability

Naïve Bayesian Classifier

- Naïve Bayesian classifier calculate this posterior probability using Bayes’ theorem, which is as follows.

- From Bayes’ theorem on conditional probability, we have

where,

(Y)

Note:

- is called the evidence (also the total probability) and it is a constant.
- The probability *P(Y|X)* (also called class conditional probability) is therefore proportional to *P(X|Y)*.
- Thus, *P(Y|X)* can be taken as a measure of *Y* given that *X*.

$P(Y|X)$

Naïve Bayesian Classifier

- Suppose, for a given instance of *X* (say *x* = () and ….. .

- There are any two class conditional probabilities namely *P(Y|X=x)* and *P(YX=x)*.

- If *P(YX=x)* > *P(YX=x)*, then we say that is more stronger than for the instance *X = x*.

- The strongest is the classification for the instance *X* = *x*.

Naïve Bayesian Classifier

- **Example:** With reference to the Air Traffic Dataset mentioned earlier, let us tabulate all the posterior and prior probabilities as shown below.

		Class			
Attribute		On Time	Late	Very Late	Cancelled
Day	Weekday	9/14 = 0.64	½ = 0.5	3/3 = 1	0/1 = 0
Season					
	Winter	2/14 = 0.14	2/2 = 1	2/3 = 0.67	0/1 = 0

Naïve Bayesian Classifier

		Class			
Attribute		On Time	Late	Very Late	Cancelled
Fog					
	High	4/14 = 0.29	1/2 = 0.5	1/3 = 0.33	1/1 = 1
Rain					
	Heavy	1/14 = 0.07	1/2 = 0.5	2/3 = 0.67	1/1 = 1
Prior Probability					

Naïve Bayesian Classifier

Instance:

Week Day	**Winter**	**High**	**Heavy**	**???**

Case1: Class = On Time : 0.70 × 0.64 × 0.14 × 0.29 × 0.07 = 0.0013

Case2: Class = Late : 0.10 × 0.50 × 1.0 × 0.50 × 0.50 = 0.0125

Case3: Class = Very Late : 0.15 × 1.0 × 0.67 × 0.33 × 0.67 = 0.0222

Case4: Class = Cancelled : 0.05 × 0.0 × 0.0 × 1.0 × 1.0 = 0.0000

Case3 is the strongest; Hence correct classification is **Very Late**

Naïve Bayesian Classifier

Algorithm: Naïve Bayesian Classification

Input: Given a set of *k* mutually exclusive and exhaustive classes *C* = , which have prior probabilities $P(C_1)$, $P(C_2)$,….. $P(C_k)$.

There are *n*-attribute set *A* = which for a given instance have values = , = ,….., =

Step: For each , calculate the class condition probabilities, *i* = 1,2,…..,*k*

Output: is the classification

Note: , because they are not probabilities rather proportion values (to posterior probabilities)

Naïve Bayesian Classifier

Pros and Cons

- The Naïve Bayes' approach is a very popular one, which often works well.
- However, it has a number of potential problems
  - It relies on all attributes being categorical.
  - If the data is less, then it estimates poorly.

Naïve Bayesian Classifier

Approach to overcome the limitations in Naïve Bayesian Classification

- Estimating the posterior probabilities for continuous attributes

  - In real life situation, all attributes are not necessarily be categorical, In fact, there is a mix of both categorical and continuous attributes.

  - In the following, we discuss the schemes to deal with continuous attributes in Bayesian classifier.

1. We can discretize each continuous attributes and then replace the continuous values with its corresponding discrete intervals.

2. We can assume a certain form of probability distribution for the continuous variable and estimate the parameters of the distribution using the training data. A Gaussian distribution is usually chosen to represent the posterior probabilities for continuous attributes. A general form of Gaussian distribution will look like

   where, denote mean and variance, respectively.

Naïve Bayesian Classifier

For each class C_i, the posterior probabilities for attribute A_j (it is the numeric attribute) can be calculated following Gaussian normal distribution as follows.

Here, the parameter can be calculated based on the sample mean of attribute value of for the training records that belong to the class .

Similarly, can be estimated from the calculation of variance of such training records.

Naïve Bayesian Classifier

M-estimate of Conditional Probability

- The M-estimation is to deal with the potential problem of Naïve Bayesian Classifier when training data size is too poor.
    - If the posterior probability for one of the attribute is zero, then the overall class-conditional probability for the class vanishes.
    - In other words, if training data do not cover many of the attribute values, then we may not be able to classify some of the test records.
- This problem can be addressed by using the M-estimate approach.

M-estimate Approach

- M-estimate approach can be stated as follows

where, *n* = total number of instances from class

= number of training examples from class that take the value

m = it is a parameter known as the equivalent sample size, and

p = is a user specified parameter.

Note:

If *n* = 0, that is, if there is no training set available, then = *p*,

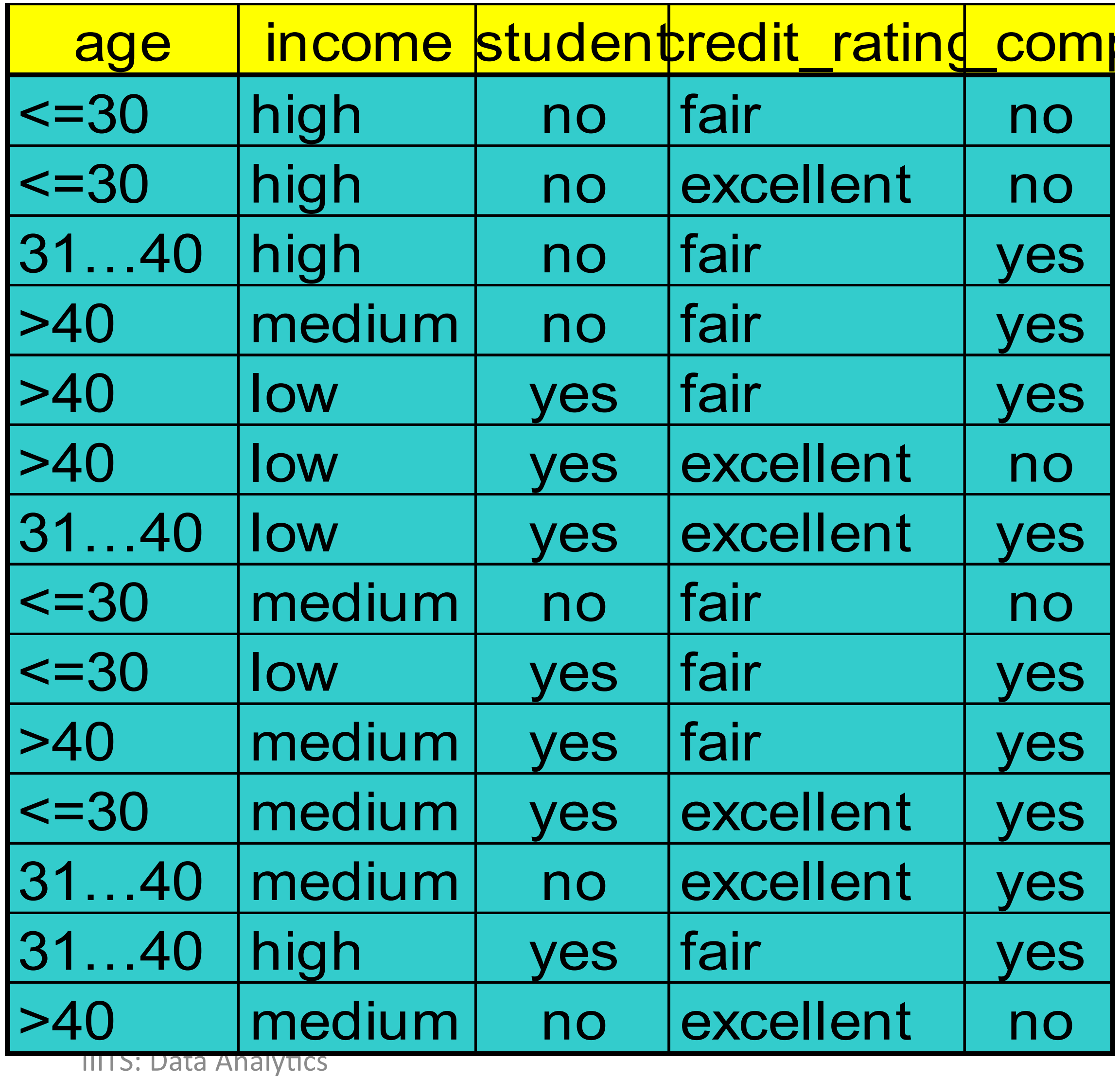
so, this is a different value, in absence of sample value.

A Practice Example

Example 8.4

Class:
C1:buys_computer = ‘yes’
C2:buys_computer = ‘no’

Data instance
X = (age <=30,
Income = medium,
Student = yes
Credit_rating = fair)



age	income	student	credit_rating	comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

A Practice Example

- P(C_i): P(buys_computer = “yes”) = 9/14 = 0.643
  P(buys_computer = “no”) = 5/14= 0.357

- Compute P(X|C_i) for each class
  P(age = “<=30” | buys_computer = “yes”) = 2/9 = 0.222
  P(age = “<= 30” | buys_computer = “no”) = 3/5 = 0.6
  P(income = “medium” | buys_computer = “yes”) = 4/9 = 0.444
  P(income = “medium” | buys_computer = “no”) = 2/5 = 0.4
  P(student = “yes” | buys_computer = “yes) = 6/9 = 0.667
  P(student = “yes” | buys_computer = “no”) = 1/5 = 0.2
  P(credit_rating = “fair” | buys_computer = “yes”) = 6/9 = 0.667
  P(credit_rating = “fair” | buys_computer = “no”) = 2/5 = 0.4

- **X = (age <= 30 , income = medium, student = yes, credit_rating = fair)**

P(X|C_i) : P(X|buys_computer = “yes”) = 0.222 × 0.444 × 0.667 × 0.667 = 0.044
P(X|buys_computer = “no”) = 0.6 × 0.4 × 0.2 × 0.4 = 0.019

P(X|C_i)*P(C_i) : P(X|buys_computer = “yes”) * P(buys_computer = “yes”) = 0.028
P(X|buys_computer = “no”) * P(buys_computer = “no”) = 0.007

Therefore, X belongs to class (“buys_computer = yes”)

Thank You!